IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 18, 2018, accepted October 3, 2018, date of publication October 15, 2018, date of current version November 9, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2876122

Transfer Learning for Wearable Long-Term Social Speech Evaluations

YUANPENG CHEN[1], BIN GAO[1], LONG JIANG[1], KAI YIN[2], JUN GU[1], AND WAI LOK WOO[3]
[1]School of Automation Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China
[2]Research Institute Electronic Science and Technology, University of Electronic Science and Technology of China, Chengdu 611731, China
[3]School of Engineering, Newcastle University, Newcastle upon Tyne NE1 7RU, U.K.

Corresponding author: Bin Gao (bin_gao@uestc.edu.cn)

ABSTRACT With an increase of stress in work and study environments, mental health issue has become a major subject in current social interaction research. Generally, researchers analyze psychological health states by using the social perception behavior. Speech signal processing is an important research direction, as it can objectively assess the mental health of a person from social sensing through the extraction and analysis of speech features. In this paper, a series of four-week long-term social monitoring experiment study using the proposed wearable device has been conducted. A set of well-being questionnaires among of a group of students is employed to objectively generate a relationship between physical and mental health with segmented speech-social features in completely natural daily situation. In particular, we have developed transfer learning for acoustic classification. By training the model on the TUT Acoustic Scenes 2017 data set, the model learns the basic scene features. Through transfer learning, the model is transferred to the audio segmentation process using only four wearable speech-social features (energy, entropy, brightness, and formant). The obtained results have shown promising results in classifying various acoustic scenes in unconstrained and natural situations using the wearable long-term speech-social data set.

INDEX TERMS Long-term social monitoring, psychological, transfer learning.

I. INTRODUCTION

Natural, efficacious, and trustworthy long-term social-speech monitoring by using wearable device is required for researchers to find and establish the interrelationships of Social Signal Processing (SSP) and Physical Mental Health (PMH). Developing audio processing method for extracting social-audio features are just as important as conscious content for evaluating human wellbeing. Psychologists speculate these features have evolved as a way to establish hierarchy and group cohesion because they function as a subconscious discussion for relationships, personality, mental health and etc. The standard methods [1], [2] to measure and evaluate social behavior include: i) using surveys for mental health evaluation where it often suffers from subjectivity and memory effects, ii) using cameras to capture behavioral responses but this is extremely expensive and the range of measurement is limited within a particular place. Thus, SSP [3], [4] is an active research field where Wearable Intelligent Devices [5] (WID) can objectively sense and understand human wellbeing. In addition, SSP has already attracted researchers in such as psychology, intelligent management and healthcare.

The wearable social computing system such as Sociometrical Badge [6] and Wearable Audio Meter (WAM) [7] encounters an ability of capturing social signals in persuasive manner. However, the Sociometrical Badge measures only the number of face-to-face interactions and the conversational time. The WAM device retains the original speech file where this suffers from the privacy issues. In addition, the requirement of the storage space for huge speech data limits the long-term use of the wearable devices.

Speaker segmentation is a crucial part for wearable SSP analysis. High-efficiency segmentation method will provide more accurate extraction of the speech social features. This leads to the better understanding of the evaluation of the social state. The segmentation methods mainly concentrated in three aspects: supervised, unsupervised segmentation algorithms and semi-supervised learning. The supervised segmentation algorithm requires in advance the label of the training data. Text annotations are provided as labels in the field of speech segmentation. Barchiesi *et al.* [8] describe classifiers and low-level time and frequency features in speech segmentation.

Typical audio features consist of mel-frequency cepstral coefficients and spectral features such as fundamental frequency, frequency components, spectral centroid spectral flux, and etc. Recently, the histograms of sound events and gradients learned from time-frequency representation are used as audio features for segmentation. In addition, model based methods are investigated for segmentation. These include Hidden Markov Models (HMMs) [9], gaussian mixture models (GMMs) [10], support vector machines (SVM) [11] and nonnegative matrix factorization (NMF) [12]–[14]. Notwithstanding above, the recent rise of the deep learning model is developed in the field of speech segmentation. It consist of using convolutional neural networks (CNN) [15] and long short-term memory (LSTM) [16], [17]. Compared with the classic statistical model, the deep learning model can learn wider features and obtain relatively better results. However, the supervised segmentation methods often require a larger dataset, human annotation and higher time cost [18].

As the size of datasets continues to grow, annotation work is increasingly time-consuming and expensive. Thus, unsupervised learning segmentation methods have the potential to ameliorate the problem where these models do not require the labeled annotations. Sharma and Mammone [19] proposed a blind speech segmentation method without linguistic knowledge. Bridle and Sedgwick [20] presented an automatic method of dividing the pattern from one utterance of a word into a sequence of segments. Estevan *et al.* [21] presented an unsupervised speech segmentation method based on maximum margsin clustering. Qiao and Shimomura [22] tried to solve the segmentation problem by searching for the ''optimal segmentation'' by using a probabilistic framework. Dusan and Rabiner [23] presented a method that detects boundaries by searching for the peaks in a spectral transition measure. Unsupervised HMMs [24]–[26] and neural networks (NNs) [27]–[29] have been proposed for speech segmentation. Xu *et al.* [30] studied maximum margin clustering (MMC) for speech segmentation. In recent study, Wang *et al.* [31] proposed DNN-based acoustic features to learn data from external set of unlabeled data through unsupervised learning. In [32], Kamper *et al.* developed cAE-learned features to improve the segmentation performance. Stacked autoencoders (AEs) has been used in segmenting audio data [33]–[35]. Several researchers have used weak top-down supervision in training unsupervised NN-based models [27], [29], [36]. Expectation-maximization based semi-supervised learning has been studied for various acoustic pattern recognition problems such as speaker identification [37] and musical instrument recognition [38].

In recent years, transfer learning [39] is applied to solve the problem which lacks sufficient data in machine learning. It is mainly used to solve the problem of covariate shift [40] or dataset deviation [41], [42] in traditional machine learning. These applications are text analysis [43], natural Language processing [44] and image classification [45]. Transfer learning has been successfully applied in the field of computer vision [46]. There are a large number of public databases in the field of computer vision. For example ImageNet, COCO and other datasets. Pre-trained networks on large data sets have become a standard. In transfer learning, the model can learn sufficient information in largescale public databases, and thus can achieve good results in transferring to other tasks. In the audio field, the application of transfer learning is not as widespread as in the field of computer vision. However, it has made a few considerable achievements. In the field of emotional audio research, it is possible to transfer learning in the same-domain dataset [47] even between music and speech domains [48].

However, both current supervised and unsupervised methods have challenges to achieve higher segmentation performance since only limited audio features are provided. As for long-term social-speech monitoring, the limited features are embedded in wearable social-sensing computing system due to the computation constraint and power saving. In this paper, a novel long time speech segmentation system for the evaluation of SSP is proposed. The contributions are summarized as follows.

(1) Designing a wearable intelligent device to evaluate mental health based on long term speech-social feature analysis which can deal with speech signals in persuasive manner. In order to unobtrusively and continuously marking the granular details of behaviors and contexts, limited four speech-social feature signals are used for analysis and embedded to avoid directly recording their voice. We designed an experiment for university students to collect data for a long period of one month for evaluating their physical and mental health.
(2) The parameter/model transfer learning is applied in the field of acoustic classification and its effect is studied through a model-based transfer learning algorithm. The transfer learning model is verified to solve the problem of insufficient training samples as it often happens in the traditional model.
(3) Finally, the interrelationships between the segmented speech-social features with anxiety level state are established. The results are used to assist in studying the mental health of college students.

The paper is organized as follows: Section II introduces the main framework of long-term speech social scene analysis and the transfer learning algorithms. Experimental results and a series of performance comparisons and analysis are described in Section III. Finally, Section IV concludes the paper.

II. PROPOSED SYSTEM DESCRIPTION

A. LONG-TERM SPEECH SOCIAL SCENE ANALYSIS FRAMEWORK

In this section, the basic framework of the speech social scene analysis is described. Fig.1 includes a system block diagram of the long-time speech segmentation and transfer learning as well as long-term wearable speech-social scene analysis. Specifically, the system diagram describes how model-based

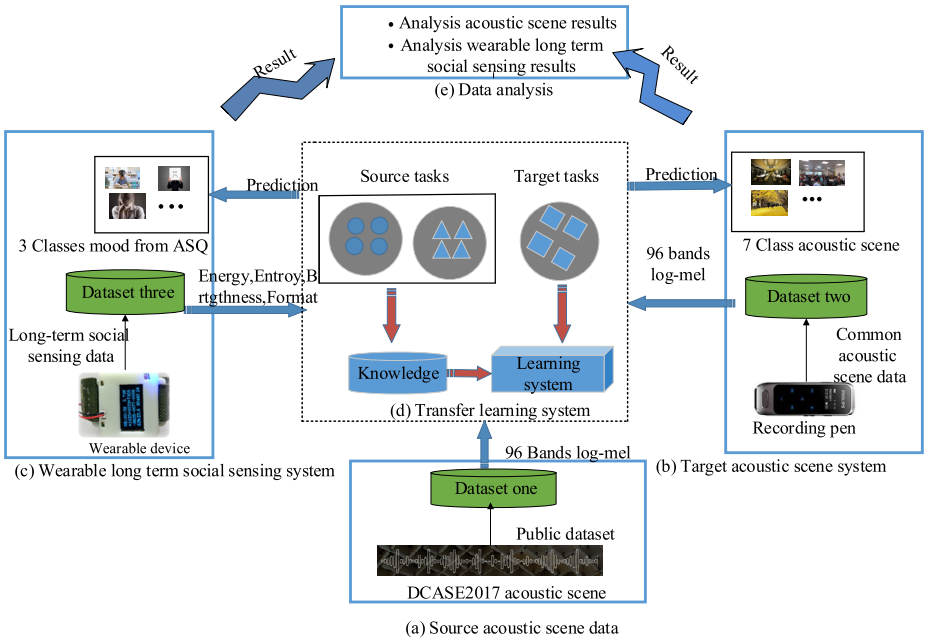


FIGURE 1. System block diagram of long-time speech segmentation. (a) Source acoustic scene data. (b) Source acoustic scene data. (c) Wearable long term social sensing system. (d) Transfer learning system. (d) Transfer learning system.

transfer learning can be used in acoustic scenarios with limited audio features.

As shown in Figure 1, the overall system consists of five sub-systems: a) Source acoustic scene data, b) Target acoustic scene system, c) Target wearable long-term social sensing system, d) Transfer learning system, e) Data analysis. DCASE2017 acoustic data serves as the training data set. Two testing datasets are involved where they are target acoustic scenes system and the target wearable long-term speech-social scene analysis. The dataset in the target acoustic scene system is the common acoustic scene collected by the audio recording pen. The data in the target wearable long-term social sensor system comes from the designed wearable wristband device. The model is trained on the source dataset through a model-based transfer learning algorithm and applied it to two target testing datasets to analyze the results of acoustic scene classification and wearable long-term social sensing results.

B. DESCRIPTION FEATURES AND WEARABLE SOCIAL-SENSING PLATFORM

All datasets are cut into non-overlapping sequences of 10s equal length which is equivalent to only training on a small ''sliding'' window instead of training the entire sequence. The features of the three datasets are summarized in Table 1.

In dataset one and dataset two, we uses the FreesoundExtractor4 to extract features from the audio analysis software Essentia Open Source Library. For the selection of features, the relevant detailed description can be found in [49].

The network input for these two datasets is the 96 bands log-mel energies. Firstly, the Short-Time Fourier Transform (STFT) of 2048 frame size and 1024 hop size is calculated. The 96 bands Mel energies are calculated by using the SIFT spectrogram. The formula for calculating Mel energy is log (10000 x Mel + 1). The operations are all conducted by using the Essentia open-source library. For dataset three, energy, entropy, brightness and formant features are extracted from the wearable device [50]–[52].

TABLE 1. Summary of features in dataset.

Dataset	Features
Dataset one	96 Bands log-mel
Dataset two	96 Bands log-mel
Dataset three	Energy,Entropy,Brightness,Format

We asked participants to fill out a series of questionnaires to assess their mental health. The rosenberg self-esteen (RESE) [53] is a self-esteem measurement tool. Positive and negative affect schedule (PANAS) [54] is used to assess the cognitive and well-being of the test. It is composed of 20 self-reported questions. The state-trait anxiety inventory (STAI) [55] can measure both state anxiety and trait anxiety, consisting of 40 self-reported issues. The neuroticism extraversion openness to experience five-factor inventory (NEO-FFI) [56] consists of 60 personality traits. The beck depression inventory (BDI) [57] is used to measure the subject's degree of depression. The autism-spectrum quotient (ASQ) [52] is a self-administered questionnaire for the screening of autistic symptoms in healthy subjects and individuals with highly functional ASD with a good predictive validity.

FIGURE 2. Wearable device wear mode and hardware platform.

These questionnaires mainly include two aspects of personality and emotional state.

Figures 2 shows the porotype of the wearable device and the related hardware platform. The microprocessor is ARM-Conrtex4 with DSP function and the model is STM32F405. The sensor unit contains activity sensors (MPU6050), environmental sensors (BH750, BMP80) and body sensors (SI7021). The sampling rates are 100Hz, 0.1Hz and 0.1Hz, respectively. The audio sensor (MEMS microphone) is 8kHz. The display module is an OLED screen. The memory module is a microSD card and includes a power management unit. Specifically, the wearable device can automatically extract social related audio features and delete the original data to protect personal privacy. The wearable hardware platform will extract four socially related audio features (Energy, Entropy, Brightness, Formant). The digital signal of the audio unit is sampled and filtered using an analog to digital converter (ADC).The audio signal is filtered and amplified by an inter-integrated-circuit (I2C) bus. The collection of sensor data is controlled by threads. The purpose of the three experiments was to use the wearable social-mental-health sensing platform to analyze the relationship between the level of autism in healthy people and the long-term daily social perception features.

C. PROPOSED FRAMEWORK OF TRANSFER LEARNING SYSTEM AND METHOD

Transfer learning can be divided into four domains: Instance based transfer learning, feature based transfer learning, parameter/model-based transfer learning, relational-knowledge based transfer learning. In this paper, the parameters/model-based transfer learning technology is used and adapted for the proposed investigation. Parameters/model-based transfer learning is originated from the source domains and destinations. Model-based transfer learning refers to the method of finding the parameter information shared between them from the source domain and the target domain to implement the transfer. The assumption required by this transfer method is that the data in both source domain and target domain can share the parameters of the models. Yosinski *et al.* [59] conducted a study of the mobility of deep neural networks. The goal is to study the mobility of transfer-learning deep networks. Zhao *et al.* [60] proposed the TransEMDT method the existing labeled data and the decision tree is used to construct a robust behavior recognition model. Then, for the uncalibrated data, the K-Means clustering method is used to find the optimal calibration parameters. Wei *et al.* [61] added social information to the regular term of the transfer learning method and improved the method. Most of the current model-based transfer learning is combined with deep neural networks. These methods modify the structure of several existing neural networks, add domain adaptation layers to the network, and start to the joint training process. Transfer learning consists of a source task and a target task. The dataset of the source task is generally large, which allows the model to be well trained. Target tasks are generally performed by given a limited set of data. The potential useful knowledge extracted in the source task will be delivered to the target task. Let us define a labeled source domain as $\boldsymbol{D_s} = \{\boldsymbol{x_i}, \boldsymbol{y_i}\}_{i=1}^{n}$ and an unmarked target domain as $\boldsymbol{D_t} = \{\boldsymbol{y_j}\}_{j=n+1}^{n+m}$. The data distributions $\mathbf{p}(\boldsymbol{x_s})$ and $\mathbf{p}(\boldsymbol{x_t})$ of these two fields are different, that is $\mathbf{p}(\boldsymbol{x_s}) \neq p(\boldsymbol{x_t})$. The goal of the transfer learning is to learn $\boldsymbol{D_t}$ with the knowledege of $\boldsymbol{D_s}$. In this paper, we used the parameter/model-based transfer learning. Hence, the goal is to transfer the tasks of the acoustic classification and speech scene classification in the source task to specific scene classification with limited information.

The block diagram of the model-based transfer learning system is shown in Figure 3.The baseline network consists of two types of networks. One is a VGG-net (Visual Geometry Group) with a 5×5 convolution kernel, and the other is a VGG-net with a 3×3 convolution kernel. The system includes source/target dataset, speech feature, parameter/model learning. The audio feature is composed of features from three datasets. The overall system block diagram includes a total of three stages. The first stage is the training of the baseline VGG-net on the source dataset. The second stage is based on the transfer process of the VGG-net (5×5 convolutional) model. B8A represents the weight of the frozen 8 layer from the A network, and then fine-tuning the B network. The third stage is based on the transfer process of the VGG-net (3×3 convolutional) model. C3A represents the weight of our frozen 3-layer from the Network A, and then fine-tuning Network C.

1) VGG NETWORK FRAMEWORK

As the VGG network is used [62] to adapt to our dataset, we changed the structure of several specific networks. We use two 3×3convolution kernel instead of one 5×5 convolution kernel as shown in Figure 4.The theoretical basis is that the size of the receptive field of two 3×3 convolution kernels and one 5×5 convolution kernel is the same [63].

The receptive field is calculated as

$$r_{out} = r_{in} + (k - 1) * j_{in} \tag{1}$$

where r_{out} is the receptive field of the current layer, r_{in} is the input receptive field, k is the size of the current layer core, the convolutional layer is the size of the convolution kernel,

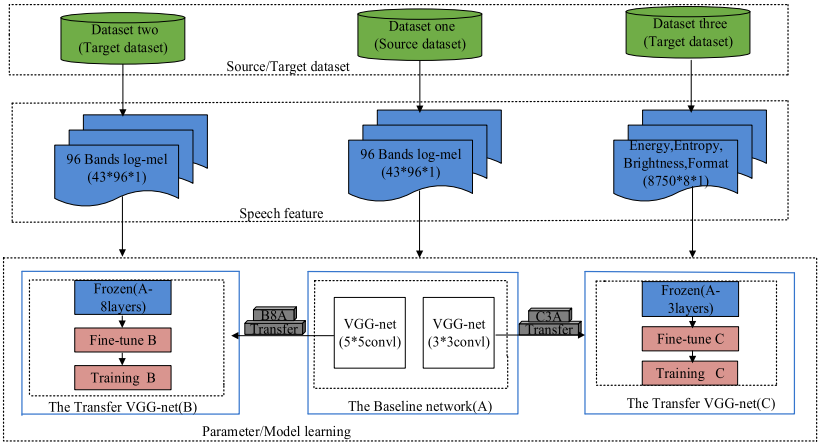


FIGURE 3. System block diagram of transfer learning.

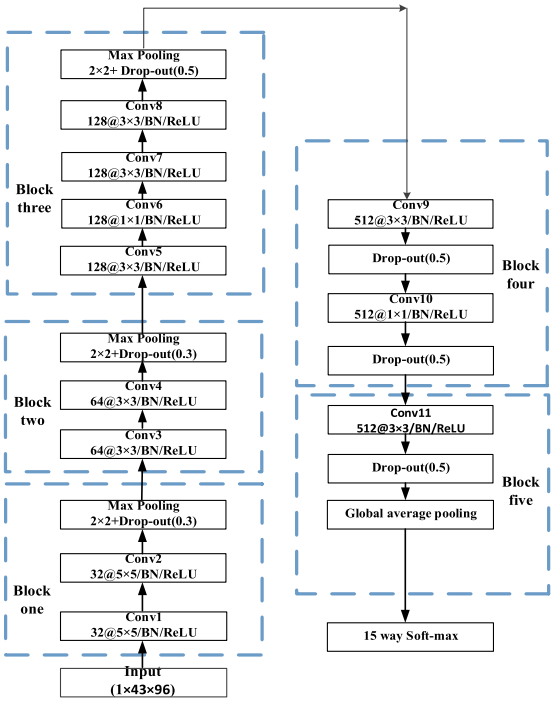


FIGURE 4. VGG-net network structure.

and the pooled layer is the size of the pooled kernel, j_{in} is the current strides, and strides in the current layer are the product of strides in the previous layer.

After a simple algebraic transformation, the final expression is obtained as

$$r_n = r_{n-1} + (k_n - 1)\prod_{i=1}^{n-1} s_i \tag{2}$$

Considering the input tensor dimension of the network, we need to consider the computational complexity of the convolutional feature map size in designing the network structure. The input to the general network is a tensor of W × H × D, where W is the width, H is the height, and D is the number of channels. Different paces can be laid down horizontally and vertically.

The accuracy of the convolution output of a filter with size (width = w, height = h) and a filter with size (width = F_w, height = F_h) can be calculated by using the following expressions:

$$\text{output width} = \frac{W - F_w + 2P}{S_w} + 1 \tag{3}$$

$$\text{output height} = \frac{W - F_h + 2P}{S_h} + 1 \tag{4}$$

In deep learning open source platform Tensorflow and Keras, there are two options same and valid. In this paper, we are using same mode in which we have zero padding. Thus, the size of output will be

$$\text{output height} = \text{ceil}(\frac{H}{S_h}) \tag{5}$$

$$\text{output width} = \text{ceil}(\frac{H}{S_w}) \tag{6}$$

The VGG-net is the current state-of-the-art deep convolutive neural network and has the advantage of stacking multiple small convolution kernels without using pooling operation where not only it can increase the characterization depth of the network but also limit the number of parameters. For example, by stacking three 3×3 convolution layers instead of using a single 7×7 layer, several limitations can be overcome. Firstly, this combines three nonlinear functions as it makes the decision function more discriminative and characteristics. Secondly, the amount of parameters is reduced by 81% while the receptive field remains unchanged. In addition, the use of small convolution kernels plays the role of a regularizer and increases the effectiveness of different convolution kernels. It leads to strong expansion, good generalization of migration, and simple structure. The entire network uses the same convolution size and maximum pool size. At the same time, as the model architecture is wider, the increase in the calculation amount is slowed down.

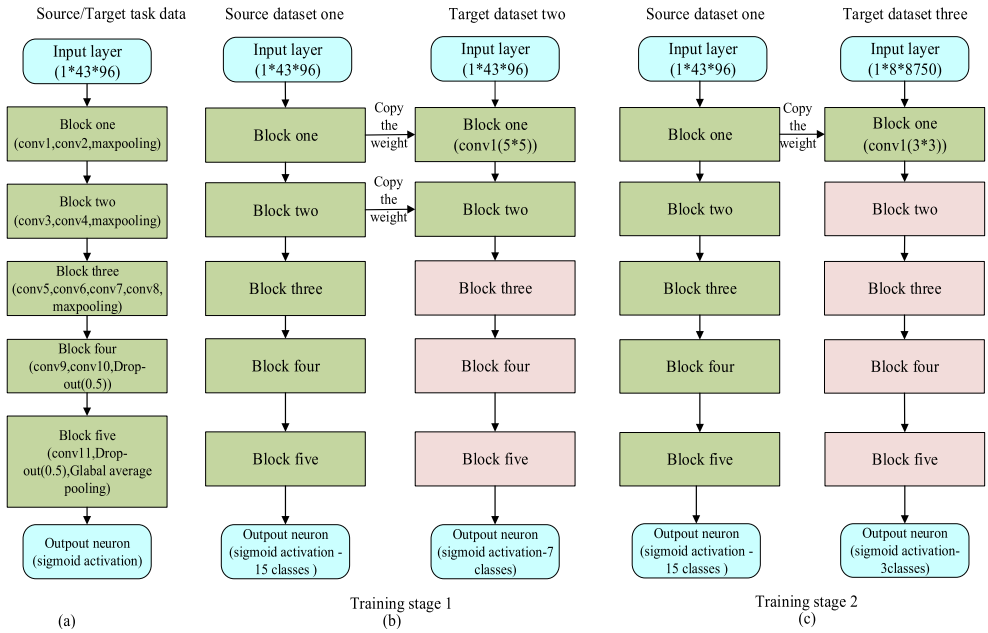


FIGURE 5. Model-based transfer process. (a) The baseline network without transmission is implemented via network. (b) The transfer process where Dataset one is the source dataset and Dataset two is the target dataset. (c) The transfer process where Dataset one is the source dataset and Dataset three is the target dataset.

2) MODEL BASED TRANSFER LEARNING

We use a model-based VGG-net to transfer within the acoustic classification field. In Figure 5 there are three parts: in the first part is a tensor representation after the feature extraction, and the second part is the network topic component composed of convolution, pooling, and drop-out. The third part is the output module which consists of an output sigmoid, which is used to predict whether the sound event sequence is detected in the output audio frame. The shallow structure generally extracts simple features, and the deep structure extracts more abstract features based on the features acquired by the shallow structures. The basic architecture of the network was used for Training source tasks (training stage 1). The low-level features obtained at this stage are fixed and only the higher-level features of the target dataset (training stage 2, 3) are required to be fine-tuned accordingly. In the actual operation, when the rest of the network of the target task data execution network is required to be retrained, the first and second blocks are set to be fixed.

The input size of the network is a single-channel speech feature. The network input tensor of dataset one, dataset two is [34, 96, 1] and the tensor input of network of dataset three is [8750, 8, 1]. The proposed model parameters are minimized by using random gradient descent and momentum optimization.

The initial learning rate is 0.001 and the momentum is fixed at 0.8 throughout the entire training process. All networks are trained by using a binary cross-entropy loss. Due to the scarcity of target task data, we do not use verification data due to the early stop. Instead each learning phase is terminated after 1000 periods in GTX 1080Ti.

III. EVALUATION AND ANALYSIS

A. DATASETS

The TUT Acoustic Scenes 2017 dataset is used as dataset one. It is composed of the development datasets and the evaluation audio datasets. It includes 15 acoustic scenes (Beach, Bus, Café/restaurant, Car, Citycenter, Forestpath, Grocery store, Home, Library, Metro station, Office, Park, Residential area, Train, Tram). A four-fold cross-validation is used during training. Each sound scene contains 312 recordings, each of which has a 10 seconds duration. The sampling frequency of the audio is 44.1kHz with resolution of 24 bits.

The dataset two is based on our own collection of audio conversations. It contains a dialogue database, an audio-only database. The dialogue database is composed of data from different noise scenarios. The scenes of data collection are mainly divided into Library (quiet environment), Road (noisy environment) and Restaurant (noisy-intensive environment), English recitation, Chinese recitation, English speech, Chinese speech, a total of 7 classes. Among the Library (quiet environment), Road (noisy environment) and Restaurant (noisy-intensive environment) 3 groups of data sources are grouped by eight people (four male and four female, the mother tongue is Chinese, age 20 -30 years old). We grouped eight people, male and female, male and female, female and female. The English reciation, Chinese reciation, English speech, Chinese speech data came from English classroom experiments [50]. The data includes male and female students. The sampling frequency of the audio is 44.1kHz with resolution of 24 bits.

The third dataset is composed of four audio features collected on the wearable device in a long-term collection.

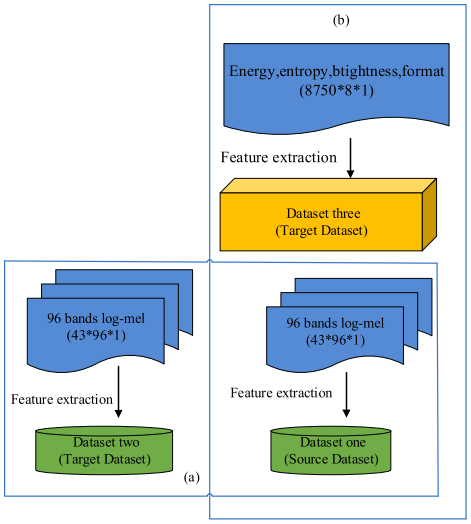


FIGURE 6. Relations between the three dataset.

We conducted a four-week long-term supervised experimental study of 16 students at the University of Electronic Science and Technology (the list of students participating in the best is provided by psychology teachers). We collect weekly questionnaires (PANAS form, SAI form, TAI form, ENO-FFI from SES from BDI-II form, etc). The wearable device is worn on the tester's commonly used wrist. The wearable device processes speech data extraction(energy, entropy, brightness, formant1, formant2, formant3, formant4, formant5) every 10 seconds. Four speech features are stored in the SD card, and no less than 8 hours of speech feature data is collected in one day. Through the wearable device, the audio data of the tester in the daily situation can be obtained. We divided the data into 3 classes through the ASQ. The selection criteria of the participants are recommended by the psychology field researchers as this study mainly concern on the students affected by psychological stress and anxiety. Participants were recruited in the college, aged 17 to 18 years. All participants were right-handed and hand normal or corrected to normal vision. Exclusion criteria included current or regular substance or medication use, current or history of medical or psychiatric disorders. They all stay at the school during this period (leave no more than 1 day, test for 28 days in total.)The sampling rate of the audio sensor is 8KHz.

Figure 6 shows the relationship between the three datasets. The dataset in (a) box is dataset one and dataset two. The extracted features are the same and the original data set is similar to the acoustic scene. The data set in (b) box is dataset one and dataset three. The extracted features are not the same and the original data set is a dissimilar acoustic scene.

B. DATA AUGMENTATION

For dataset one (DCASE2017 dataset). The data is composed of records from various sound fields. We extracted four mono-version files for each audio. Through this four monaural data, the data can be enhanced in the model. For dataset two, the same approach is taken.

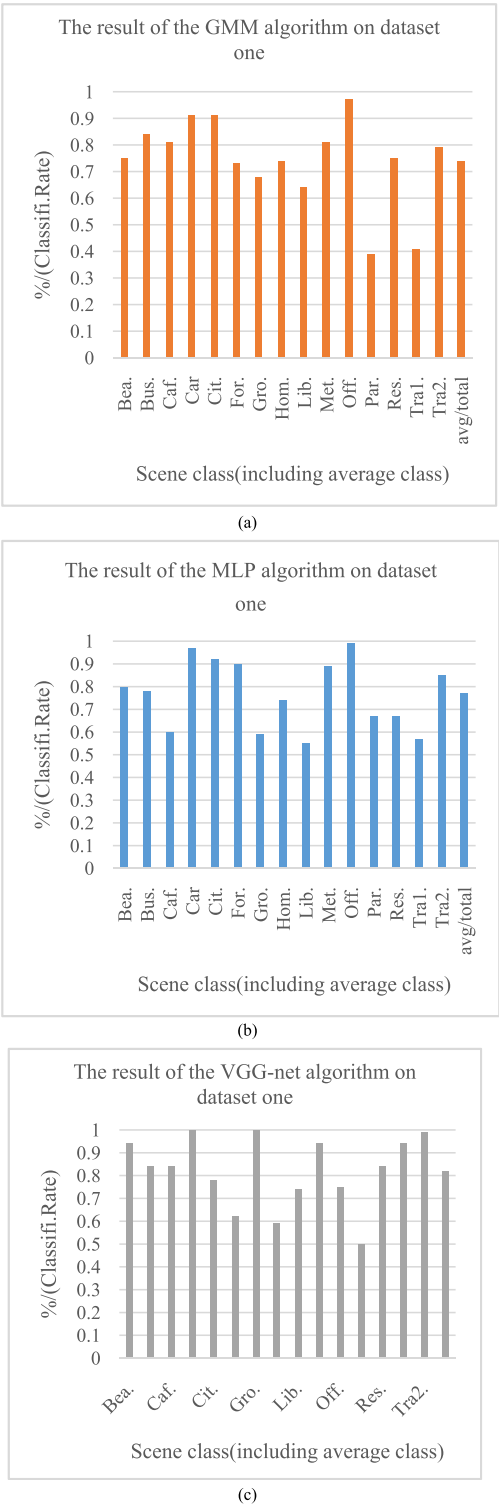


FIGURE 7. (a) The precision of the GMM on data one. (b) The precision of the MLP on dataset one. The precision of the VGG-net on dataset one.

1) Left channel
2) Right channel
3) Average channel = (left channel + right channel)/2
4) Difference channel = (left channel − right channel)/2

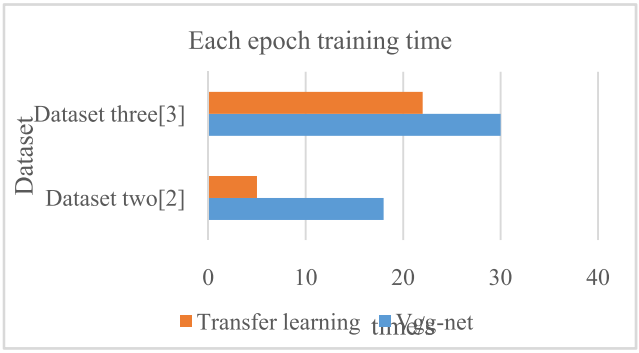


FIGURE 8. Comparison of training time.

For data augmentation, we used MWFD (multiple-width frequency-delta data segmentation) [49], [64] and batch pitch shift augmentation.

C. EXPERIMENTAL RESULTS

We use three indicators of Precision, Recall,and F1-score to measure the forecast results. Precision is the ratio of correctly predicted positive observations to the total predicted positive observations.

$$Precision = \frac{TP}{TP + FP} \tag{7}$$

Recall is the ratio of correctly predicted positive observations to the all observations in actual class - yes.

$$Recall = \frac{TP}{TP + FN} \tag{8}$$

F1-score is the weighted average of Precision and Recall Therefore, this score takes both false positives and false negatives into account. Intuitively it is not as easy to understand as accuracy, but F1 is usually more useful than accuracy, especially if you have an uneven class distribution.

$$F1 - Score = 2 * \frac{Recall * Precision}{(Recall + Precision)} \tag{9}$$

The results of the three algorithms are compared in Figure 7(a)-(c). These are gaussian mixture model (GMM), multilayer perceptron (MLP), and VGG-net. A total of 15 Acoustic Scene classes are composed of Beach, Bus, Café/restaurant, Car, City center, Forest path, Grocery store, Home, Library, Metro station, Office, Park, Residential area, Train (Tra1.), Tram (Tra2.). From the results of avg/total, it can be seen that the VGG-net algorithm is more accurate than the GMM algorithm and the MLP algorithm. The overall performance at each class of VGG-net algorithm is also relatively better. Therefore, we chose the VGG-net as the proposed reference network for transfer learning.

Notwithstanding above, we have compared the proposed scheme with another state-of-the-art learning algorithm MResNnet-34 [65].

In the deep residual convolutional neural network architecture (MResNet-34), there are two type of residual blocks: A and B. The difference between block A and block B is the starting point of the shortcut connection. In block A, the shortcut connection starts after the batch normalization and the rectified linear unit activation. I n the block B, the shortcut connection starts directly from the input of the block. The results of MResNet-34 and VGG-net are shown in the following table 2.

TABLE 2. Pecision resule of the Vgg-net and MrseNet-34 on Dcase2017 dataset.

Scene	VGG-net	MResNet-34
Bea.	94%	85.3%
Bus	84%	90.1%
Caf.	84%	75.0%
Car	100%	97.1%
Cit.	78%	90.7%
For.	62%	93.3%
Gro.	100%	83.0%
Hom.	59%	82.7%
Lib.	74%	73.4%
Met.	94%	98.1%
Offi.	75%	99.0%
Par.	50%	70.2%
Res.	84%	81.4%
Tra1.	94%	75.6%
Tra2.	99%	90.4%
Avg/total	82%	85.4%

From the results of the table, it can be seen that the VGG-net algorithm and the MResNet-34 algorithm have overall similar results. Compared to VGG-net, the MResNet-34 network layer is deeper. It has the advantage of using the skip connection. In addition, due to the fewer wearable audio features can be used, a model-based transfer learning with a shallower layer is more suitable. We implemented these algorithms on the GTX1080Ti platform.

We compared the knowledge transfer learning algorithm [66] with VGG-net. The N_s^{III} algorithm is based on the transfer of knowledge. The algorithm uses the CNN model to identify the audio in the target task, these methods to transfer knowledge are also helpful in higher level semantic understanding. The precision results are shown in the table 3.

Through the Table 3, a knowledge-based transfer learning algorithm and VGG-net are validated. It is seen that the performance of knowledge-based transfer learning algorithms is not superior to the VGG-net algorithm. The main reason is that the relational-knowledge based transfer learning algorithm depends on the degree of similarity between the

TABLE 3. Pecision resule of the Vgg-net and N_s^{III} on Dcase2017 dataset.

Scene	VGG-net	N_s^{III}
Bea.	94%	71.9%
Bus	84%	82.4%
Caf.	84%	73.8%
Car	100%	89.9%
Cit.	78%	93.3%
For.	62%	97.4%
Gro.	100%	84.6%
Hom.	59%	69.4%
Lib.	74%	73.6%
Met.	94%	80.2%
Off.	75%	85.1%
Par.	50%	46.9%
Res.	84%	63.9%
Tra1.	94%	52.3%
Tra2.	99%	84.0%
Avg/total	82%	76.6%

TABLE 4. Evaluation of results in the common acoustic scene dataset.

Methods	Classes	Evaluation Index		
		Precision	Recall	F1-score
VGG-net	Lib.	0.59	0.61	0.60
	Res.	0.49	0.64	0.55
	Roa.	1.00	0.19	0.32
	EnR.	0.78	1.00	0.87
	ChR.	0.72	1.00	0.84
	EnS.	0.84	0.70	0.76
	ChS	1.00	0.70	0.82
	Avg/total	0.78	0.73	0.70
Model-based transfer learning	Lib.	0.71	0.83	0.77
	Res.	0.70	0.42	0.53
	Roa.	0.50	0.25	0.33
	Enr.	0.55	0.84	0.67
	Chr.	0.74	0.98	0.84
	Ens.	0.89	0.74	0.81
	Chs.	0.94	0.65	0.77
	Avg/total	0.72	0.70	0.69

source domain and the target domain. Model-based transfer learning algorithms can make full use of the similarities between models.

Table 4 shows the results of the Precision, Recall, F1-score based on the common acoustic scene data transfer learning algorithm and VGG-net algorithm, respectively. In comparison, it can be seen that the accuracy of the transfer learning is similar to that of the VGG-net algorithm. There are 7 classes, these are Library, Restaurant, Road, English recitation, Chinese recitation, English speech, Chinese speech. It should be noted that dataset one is purely acoustic and the scene class does not contain speakers. In the Dataset two, scene class contains a large number of speakers. In particular, the dataset two scene contains Chinese speech and Chinese recitation.

Among the six scene classes, Chinese Speech has a high accuracy rate of 90%. Recall and F1-score are higher in the Chinese recitation class. The road class performed poorly in the three indicators. Among the three scenario classes of Library, Restaurant, Road are worse than those of English recitation, Chinese recitation, English speech, Chinese speech, respectively. The reason is that there is considerable more noise in the three classes and there is almost no noise in the transfer weights, which affects the generalization of the model.

In Table 5, The results of model-based transfer learning and VGGs-net on this dataset are not significantly different. The source domain of the transfer learning algorithm is from The DCASE2017 Acoustic Scene dataset and the target domain is the long-term special sensing dataset, with classification index based on the ASQ. The score segment is between 10-20 for the low ASQ class, the 20-30 fraction for the middle ASQ class, and the 30-40 fractional segment high ASQ class. A limited number of 8 features: energy, entropy, brightness, formant1, formant2, formant3, formant4, formant5 and ASQ have a certain relationship. Both algorithms have better predictions for high ASQ population accuracy. The transfer-based learning can accelerate the training of the model under the premise of ensuring accuracy. We will discuss this issue in the next summary.

D. ANALYSIS AND DISCUSSION

In Figure 8, it can be seen that a substantial advantage of transfer learning is the computational savings of the training resource. It illustrates the training time for each epoch. The reason for the difference in the training time between the two kinds of dataset and VGG-net is that the number of transfer layers is different. In general, the greater the number of transfer layers, the shorter the training time

Based on the above discussion, the four wearable embedded speech-social features have been shown to have a direct relationship with the mental health status of a subject. Through the prediction results, we can objectively analyze the relationship between physical and mental health with seg-

TABLE 5. Evaluation of results in the long-term social-sensing dataset.

Methods	Classes	Evaluation Index		
		Precision	Recall	F1
VGG-net	Low (10-20 of ASQ)	0.68	0.57	0.62
	Medium (20-30 of ASQ)	0.82	0.65	0.73
	High (30-40 of ASQ)	0.75	0.87	0.81
	Avg/total	0.76	0.72	0.74
Less-features transfer learning	Low (10-20 of ASQ)	0.82	0.70	0.74
	Medium (20-30 of ASQ)	0.76	0.57	0.65
	High (30-40of ASQ)	0.69	0.69	0.69
	Avg/total	0.75	0.64	0.69

mented speech-social features in completely natural daily situation. Predictions on common acoustic scene data are higher than the long-term social sensing data. This is attributed to the rich information contained in the long-term social perception dataset and the transfer learning obtained from previous training.

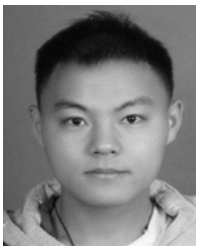
IV. CONCLUSION

This paper proposes a portable wearable device to perform long-term health monitoring. In particular, four wearable embedded speech-social features have been identified and used for mental health assessments where we can avoid privacy issues and reduce computational load. Through a four-week long experimental data, the interrelationship between the wearable captured segmented phonetic social features and the features of ASQ has been objectively established. In the least, it is shown that subjects with ASQ can be monitored for mental health using only the four speech-social features. In addition, we have successfully verified the application of a model-based model transfer learning algorithm using long-term social perception datasets.

REFERENCES

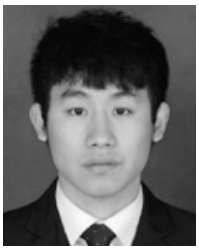
[1] A. Pentland and A. Esposito, "Secret signals," *Nature*, vol. 457, pp. 528–530, Jan. 2009.
[2] A. Pentland, "To signal is human," *Amer. Sci.*, vol. 98, no. 3, pp. 203–211, 2010.
[3] A. Pentland, "Socially aware, computation and communication," *IEEE Comput.*, vol. 38, no. 3, pp. 33–40, Mar. 2005.
[4] A. Pentland, "Social Signal Processing," *IEEE Signal Process. Mag.*, vol. 24, no. 4, pp. 108–111, Jul. 2007.
[5] A. Pentland, "Wearable intelligence," in *Scientific American Presents*. 1998, pp. 90–95.
[6] O. Olguín and S. Badges, *Wearable Technology for Measuring Human Behavior*, Cambridge, MA, USA: MIT Press, 2007.
[7] B. Gao and W. L. Woo, "Wearable audio monitoring: Content-based processing methodology and implementation," *IEEE Trans. Human-Mach. Syst.*, vol. 44, no. 2, pp. 222–233, Apr. 2014.
[8] D. Barchiesi, D. Giannoulis, D. Stowell, and M. D. Plumbley, "Acoustic scene classification: Classifying environments from the sounds they produce," *IEEE Signal Process. Mag.*, vol. 32, no. 3, pp. 16–34, May 2015.
[9] A. J. Eronen *et al.*, "Audio-based context recognition," *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, no. 1, pp. 321–329, Jan. 2006.
[10] J.-J. Aucouturier, B. Defreville, and F. Pachet, "The bag-of-frames approach to audio pattern recognition: A sufficient model for urban soundscapes but not for polyphonic music," *J. Acoust. Soc. Amer.*, vol. 122, no. 2, pp. 881–891, 2007.
[11] A. Rakotomamonjy and G. Gasso, "Histogram of gradients of time–frequency representations for audio scene classification," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 23, no. 1, pp. 142–153, Jan. 2015.
[12] S. Innami and H. Kasai, "NMF-based environmental sound source separation using time-variant gain features," *Comput. Math. Appl.*, vol. 64, no. 5, pp. 1333–1342, 2012.
[13] J. F. Gemmeke, L. Vuegen, P. Karsmakers, B. Vanrumste, and H. Van Hamme, "An exemplar-based NMF approach to audio event detection," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust. (WASPAA)*, Oct. 2013, pp. 1–4.
[14] A. Mesaros, T. Heittola, O. Dikmen, and T. Virtanen, "Sound event detection in real life recordings using coupled matrix factorization of spectral representations and class activity annotations," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 151–155.
[15] M. Espi, M. Fujimoto, K. Kinoshita, and T. Nakatani, "Exploiting pectro-temporal locality in deep learning based acoustic event detection," *EURASIP J. Audio, Speech, Music Process.*, vol. 20, no. 1, p. 26, 2015.
[16] E. Marchi, F. Vesperini, F. Eyben, S. Squartini, and B. Schuller, "A novel approach for automatic acoustic novelty detection using a denoising autoencoder with bidirectional LSTM neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Apr. 2015, pp. 1996–2000.
[17] G. Parascandolo, H. Huttunen, and T. Virtanen, "Recurrent neural networks for polyphonic sound event detection in real life recordings," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 6440–6444.
[18] M. Valenti, A. Diment, G. Parascandolo, S. Squartini, and T. Virtanen, "DCASE 2016 acoustic scene classification using convolutional neural networks," in *Proc. Detection Classification Acoustic Scenes Events Workshop (DCASE)*, Sep. 2016, pp. 95–99.
[19] M. Sharma and R. J. Mammone, "'Blind' speech segmentation: Automatic segmentation of speech without linguistic knowledge," in *Proc. ICSLP*, Oct. 1996, pp. 1237–1240.
[20] J. Bridle and N. Sedgwick, "A method for segmenting acoustic patterns, with applications to automatic speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 1977, pp. 656–659.
[21] Y. P. Estevan, V. Wan, and O. Scharenborg, "Finding maximum margin segments in speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Honolulu, HI, USA, Apr. 2007, pp. IV-937–IV-940.
[22] Y. Qiao, N. Shimomura, and N. Minematsu, "Unsupervised optimal phoneme segmentation: Objectives, algorithm and comparisons," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2008, pp. 3989–3992.
[23] S. Dusan and L. Rabiner, "On the relation between maximum spectral transition positions and phone boundaries," in *Proc. INTERSPEECH*, 2006, pp. 17–21.
[24] M.-H. Siu, H. Gish, A. Chan, W. Belfield, and S. Lowe, "Unsupervised training of an HMM-based self-organizing unit recognizer with applications to topic classification and keyword discovery," *Comput. Speech Lang.*, vol. 28, no. 1, pp. 210–223, 2014.
[25] B. Varadarajan, S. Khudanpur, and E. Dupoux, "Unsupervised learning of acoustic sub-word units," in *Proc. ACL*, 2008, pp. 165–168.
[26] C.-Y. Lee and J. Glass, "A nonparametric Bayesian approach to acoustic model discovery," in *Proc. ACL*, 2012, pp. 40–49.
[27] G. Synnaeve, T. Schatz, and E. Dupoux, "Phonetics embedding learning with side information," in *Proc. SLT*, Dec. 2014, pp. 106–111.
[28] D. Renshaw, H. Kamper, A. Jansen, and S. J. Goldwater, "A comparison of neural network methods for unsupervised representation learning on the zero resource speech challenge," in *Proc. Interspeech*, 2015, p. 3199–3203.
[29] N. Zeghidour, G. Synnaeve, M. Versteegh, and E. Dupoux, "A deep scattering spectrum—Deep Siamese network pipeline for unsupervised acoustic modeling," in *Proc. ICASSP*, Mar. 2016, pp. 4965–4969.
[30] L. Xu, J. Neufeld, B. Larson, and D. Schuurmans, "Maximum margin clustering," in *Proc. NIPS*, Vancouver, BC, Canada, 2004, pp. 1537–1544.
[31] W. Wang, R. Arora, K. Livescu, and J. A. Bilmes, "Unsupervised learning of acoustic features via deep canonical correlation analysis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 4590–4594.

[32] H. Kamper, A. Jansen, and S. J. Goldwater. (2016). "A segmental framework for fully-unsupervised large-vocabulary speech recognition." [Online]. Available: https://arxiv.org/abs/1606.06950
[33] M. D. Zeiler *et al.*, "On rectified linear units for speech processing," in *Proc. ICASSP*, May 2013, pp. 3517–3521.
[34] L. Badino, C. Canevari, L. Fadiga, and G. Metta, "An auto-encoder based approach to unsupervised learning of subword units," in *Proc. ICASSP*, May 2011, pp. 7634–7638.
[35] L. Badino, A. Mereta, and L. Rosasco, "Discovering discrete subword units with binarized autoencoders and hidden-Markov-model encoders," in *Proc. Interspeech*, 2015, pp. 3174–3178.
[36] R. Thiolliere, E. Dunbar, G. Synnaeve, M. Versteegh, and E. Dupoux, "A hybrid dynamic time warping-deep neural network architecture for unsupervised acoustic modeling," in *Proc. Interspeech*, 2015.
[37] P. J. Moreno and S. Agarwal, "An experimental study of M-based algorithms for semi-supervised learning in audio classification," in *Proc. Int. Conf. Mach. Learn. (ICML) Workshop Continuum Labeled Unlabeled Data*, 2003.
[38] A. Diment, T. Heittola, and T. Virtanen, "Semi-supervised learning for musical instrument recognition," in *Proc. 21st Eur. Signal Process. Conf. (EUSIPCO)*, 2013, pp. 1–5.
[39] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.
[40] H. Shimodaira, "Improving predictive inference under covariate shift by weighting the log-likelihood function," *J. Stat. Planning Inference*, vol. 90, no. 2, pp. 227–244, 2000.
[41] A. Torralba and A. A. Efros, "Unbiased look at dataset bias," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 1521–1528.
[42] A. Khosla, T. Zhou, T. Malisiewicz, A. A. Efros, and A. Torralba, "Undoing the damage of dataset bias," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 158–171.
[43] R. Raina, A. Y. Ng, and D. Koller, "Constructing informative priors using transfer learning," in *Proc. 23rd Int. Conf. Mach. Learn.*, 2006, pp. 713–720.
[44] J. Blitzer, R. McDonald, and F. Pereira, "Domain adaptation with structural correspondence learning," in *Proc. Conf. Empirical Methods Natural Lang. Process. Assoc. Comput. Linguistics*, 2006, pp. 120–128.
[45] P. Wu and T. G. Dietterich, "Improving SVM accuracy by training on auxiliary data sources," in *Proc. 21st Int. Conf. Mach. Learn.*, 2004, p. 110.
[46] V. M. Patel, R. Gopalan, R. Li, and R. Chellappa, "Visual domain adaptation: A survey of recent advances," *IEEE Signal Process. Mag.* vol. 32, no. 3, pp. 53–69, May 2015.
[47] J. Deng, Z. Zhang, E. Marchi, and B. Schuller, "Sparse autoencoder-based feature transfer learning for speech emotion recognition," in *Proc. Humaine Assoc. Conf. IEEE Affect. Comput. Intell. Interact. (ACII)*, Sep. 2013, pp. 511–516.
[48] E. Coutinho, J. Deng, and B. Schuller, "Transfer learning emotion manifestation across music and speech," in *Proc. Int. Joint Conf. IEEE Neural Netw. (IJCNN)*, Jul. 2014, pp. 3592–3598.
[49] R. Gong, E. Fonseca, D. Bogdanov, O. Slizovskaia, E. Gomez, and X. Serra, "Acoustic scene classification by fusing LightGBM and VGG-net multichannel predictions," in *Proc. IEEE AASP Challenge Detection Classification Acoust. Scenes Events*, 2017, pp. 1–4.
[50] J. Gu *et al.*, "Wearable social sensing: Content-based processing methodology and implementation," *IEEE Sensors J.*, vol. 17, no. 21, pp. 7167–7176, Nov. 2017.
[51] D. O. Olguin, B. N. Waber, T. Kim, A. Mohan, K. Ara, and A. Pentland, "Sensible organizations: Technology and methodology for automatically measuring organizational behavior," *IEEE Trans. Syst. Man, Cybern. B, Cybern.*, vol. 39, no. 1, pp. 43–55, Feb. 2009.
[52] S. Baron-Cohen, S. Wheelwright, R. Skinner, J. Martin, and E. Clubley, "The autism-spectrum quotient (AQ): Evidence from asperger syndrome/high-functioning autism, malesand females, scientists and mathematicians," *J. Autism Develop. Disorders*, vol. 31, no. 6, pp. 603–603, 2001.
[53] M. Rosenberg, "Self esteem and the adolescent. (Economics and the social sciences: Society and the adolescent self-image)," *New England Quart.*, vol. 148, no. 2, pp. 177–196, 1965.
[54] D. Watson, L. A. Clark, and A. Tellegen, "Development and validation of brief measures of positive and negative affect: The PANAS scales," *J. Pers. Social Psychol.*, vol. 54, no. 6, pp. 1063–1070, 1988.
[55] C. D. Spielberger, "STAI manual for the state-trait anxiety inventory," in *Self-Evaluation Questionnaire*. Palo Alto, CA, USA: Consulting Psychologitsts Press, 1970, pp. 1–24.
[56] R. R. Mccrae, *The NEO-PI/NEO-FFI Manual Supplement*. Lutz, FL, USA: Psychological Assessment Resources, 1989.
[57] M. A. Whisman, J. E. Perez, and W. Ramel, "Factor structure of the beck depression inventory—Second edition (BDI-ii) in a student sample," *J. Clin. Psychol.*, vol. 56, no. 4, pp. 545–551, 2000.
[58] S. Baron-Cohen, S. Wheelwright, R. Skinner, J. Martin, and E. Clubley, "The autism-spectrum quotient (AQ): Evidence from asperger syndrome/high-functioning autism, malesand females, scientists and mathematicians," *J. Autism Develop. Disorders*, vol. 31, no. 6, pp. 603–603, 2001.
[59] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 3320–3328.
[60] Z. Zhao, Y. Chen, J. Liu, Z. Shen, and M. Liu, "Cross-people mobile-phone based activity recognition," in *Proc. 22nd Int. Joint Conf. Artif. Intell. (IJCAI)*, vol. 11, 2011, pp. 2545–2550.
[61] Y. Wei, Y. Zhu, C. W.-K. Leung, Y. Song, and Q. Yang, "Instilling social to physical: Co-regularized heterogeneous transfer learning," in *Proc. 13th AAAI Conf. Artif. Intell.*, 2016, pp. 1338–1344
[62] H. Eghbal-Zadeh, B. Lehner, M. Dorfer, and G. Widmer, "A hybrid approach with multi-channel i-vectors and convolutional neural networks for acoustic scene classification," in *Proc. IEEE AASP Challenge Detection Classification Acoustic Scenes Events (DCASE)*, Aug. 2016, pp. 2749–2753.
[63] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. (Dec. 2015). "Rethinking the inception architecture for computer vision." [Online]. Available: https://arxiv.org/abs/1512.00567
[64] Y. Han and K. Lee. (2016). "Acoustic scene classification using convolutional neural network and multiple-width frequency-delta data augmentation." [Online]. Available: https://arxiv.org/abs/1607.02383
[65] S. Zhao, T. N. T. Nguyen, W. S. Gan, and D. L. Jones, "Acoustic scene classification using deep residual convolutional neural networks," in *Proc. IEEE AASP Challenge Detection Classification Acoust. Secenes Events (DCASE)*, 2017.
[66] A. Kumar, M. Khadkevich, and C. Fügen, "Knowledge transfer from weakly labeled audio using convolutional neural network for sound events and scenes," in *Proc. ICASSP*, Apr. 2018, pp. 326–330.



YUANPENG CHEN received the B.Sc. degree from the School of Electrical Engineering and Information Engineering, Lanzhou University of Technology, Lanzhou, China, in 2016. He is currently pursuing the M.Sc. degree in speech scene recognition and machine translation with the University of Electronic Science and Technology of China, Chengdu, China. His research interests include machine translation, image identification, speech recognition, and machine learning.

BIN GAO received the B.Sc. degree in communications and signal processing from Southwest Jiaotong University, China, in 2005, and the M.Sc. degree (Hons.) in communications and signal processing and the Ph.D. degree from Newcastle University, U.K, in 2006 and 2011, respectively. He was a Research Associate in wearable acoustic sensor technology with Newcastle University from 2011 to 2013. He is currently a Professor with the School of Automation Engineering, University of Electronic Science and Technology of China, Chengdu, China. His research interests include sensor signal processing, machine learning, social signal processing, and nondestructive testing and evaluation, where he actively publishes in these areas. He is also a very active reviewer for many international journals and long standing conferences. He has coordinated several research projects from the National Natural Science Foundation of China.



LONG JIANG received the B.Sc. degree from the School of Information Engineering, Southwest University of Science and Technology, Mianyang, China, in 2012. He is currently pursuing the M.Sc. degree in control engineering and science with the University of Electronic Science and Technology of China, Chengdu, China. His research interests include intelligent hardware and wearable sensor.

KAI YIN received the B.Sc. degree from the School of Software, Yunnan University, Kunming, China, in 2015. He is currently pursuing the M.Sc. degree in knowledge graph with the University of Electronic Science and Technology of China, Chengdu, China. His research interests include text generation and knowledge graph.

JUN GU received the B.Sc. degree from the School of Mechanical and Electronic Engineering, Nanjing Forestry University, Nanjing, China, in 2013. He is currently pursuing the M.Sc. degree in wearable device research with the University of Electronic Science and Technology of China, Chengdu, China. His research interests include embedded software development and digital signal processing.

WAI LOK WOO was born in Malaysia. He received the B.Eng. degree (Hons.) in electrical and electronics engineering and the Ph.D. degree from Newcastle University, U.K. He is currently a Senior Lecturer and the Director of Operations with the School of Electrical and Electronic Engineering, Newcastle University. His major research is in the mathematical theory and algorithms for nonlinear signal and image processing. This includes areas of machine learning for signal processing, blind source separation, multidimensional signal processing, and signal/image deconvolution and restoration. He has an extensive portfolio of relevant research supported by a variety of funding agencies. He has published over 250 papers on these topics on various journals and international conference proceedings. He received the IEE Prize and the British Scholarship to continue his research work. He is currently an associate editor of several international journals and has served as a lead editor of journals' special issues.